L4: Entry 5 of 19 File: USPT Nov 11, 2003

DOCUMENT-IDENTIFIER: US 6647383 B1
TITLE: System and method for providing interactive dialogue and iterative search functions to find information

Abstract Text (1):
A system and method for information searching comprising determination of, in fine granularity, a Community of Interest (COI), further data mining in search results, using at least one of COI and expert preferences to identify important knowledge, formulation and manipulation of results, and summarization of search results into a document like entity with dynamic attributes described. More particularly, the invention relates to a system and method for providing interactive dialogue and iterative search functions to find information on a large network of servers such as the world wide web.

Brief Summary Text (6):
Search results, as they are presented today, are not obvious to users. Also, the results are not presented in such a fashion as to take advantage of the human's ability to sift through data visually and to determine relationships among displayed objects. Most systems do not disambiguate search terminology well enough to determine what the user meant when typing a query. Typically, users do not have to perform many steps to initiate a search task; they will usually enter a few key words and then request a search. As such, there is a need to combine iterative configurable query techniques with a lexical dictionary function. This combination is currently not available in search engines.

Brief Summary Text (7):
Web search engines do not provide user access to restructure aspects of the search from a graphical user interface. When a search is conducted and results are displayed, the decisions of the search engine are not displayed such that the user can manipulate the branches and navigate down the decision and results tree, changing the attributes and thereby finding slightly different results. The user is not provided with any information on what the extent of the results may be. The user is not afforded any opportunity to reconfigure the search or the results to display the relationship among the items returned.

Brief Summary Text (8):
Some search engines provide policies that attempt to order search results based upon closeness to the query and provide metrics to the user indicating closeness. Metrics are based upon popularity, word frequency, word relationships, position of word in title or body, metatags, links within a web site, links to a page or web site, physical attributes of the web site, etc.

Brief Summary Text (15):
The results of the search should also be organized into a summary document to reveal the predictive model, sources, salient facts of the result, and links to resulting elements. This would help create content of a document about the subject search.

Brief Summary Text (16):
The results of the search should also be organized into groups, where all of the items in a group discuss similar topics. The grouping can be using information in the item (e.g., key words), or by how the item has been used in the past. The grouping can be done a number of ways; additional details will be discussed in the "Summary of the Invention" section.

Brief Summary Text (20):
A system and method for providing interactive dialogue and iterative search functions to find information among a network of servers and to display results depicting overall distribution

and relationship of results are provided. The system and method provide determination in fine granularity a Community of Interest (COI) and further evaluation of search results using COI and/or expert preferences to identify important knowledge, formulate, manipulate, and display results, and summarize search results into a document like entity with dynamic attributes. The invention is generally applicable to an information search on a large network of servers such as the world wide web where there is such a vast amount of information that it is becoming increasingly important to overcome the aforementioned difficulties in order to effectively deal with the overwhelming amount of data that a search engine might return on any given search.

Brief Summary Text (24):
Yet another advantage of the present invention is the use of new concepts in graphically presenting search results including such things as scatter grams, showing relationships among resulting elements, and the use of color, shape and other attributes to differentiate among resulting elements.

Drawing Description Text (8):
FIG. 6 is an illustration of a GUI architecture for displaying search results;

Drawing Description Text (11):
FIG. 9 is a flow chart showing how a large number of search results is handled by a Smart Search system;

Detailed Description Text (3):
COI (Community of Interest) is a policy that yields improved search results. For example, a method has been disclosed in co-pending and pending application Ser. No. 09/428,031, by Shriver and Small, filed Oct. 27, 1999, entitled METHOD FOR IMPROVING WEB SEARCHING PERFORMANCE USING COMMUNITY-BASED FILTERING, hereby incorporated by reference.

Detailed Description Text (4):
That disclosure notes that often, members of a community (an office, lab, or social organization) think about and research the same set of topics. When searching for information on the web or other electronic database, if others from one's community have recently performed the same searches, it would be helpful to know which of the results of such searches were found useful. For example, if a specific search was done by someone in the community, and after that search, a number of web pages were visited and explored, then information about that search should be useful to highlight the best page or pages from that search to another user from the same community who enters the same search.

Detailed Description Text (6):
Several methods may be used to gather the information from the searches of members of the community. In a preferred embodiment, this information gathering is done by means of examining proxy server logs. A proxy server is a server that takes requests of users and passes them to a web server, which sends requests and receives data. The web server then sends requested data to the proxy server, which transmits it back to the user.

Detailed Description Text (7):
In an alternate embodiment of the Shriver and Small method, information from the proxy server may also be examined on a real-time basis. In yet another alternate embodiment for systems without a proxy server, individual web server logs, if they exist, may be examined. These server logs contain data about the requests which have been made by users for web pages.

Detailed Description Text (9):
In a preferred embodiment of the Shriver and Small method, first the proxy server log is examined and a list of URLs visited by the user is created, ordered in temporal order. Next, the list of URLs visited is stepped through to find a search sent to a search engine. Such a search will either be found or not. If no search is found, a list of URLs visited is created for the next user. If a search is found on the list of URLs visited, the search will be replicated by sending the query to the search engine. The results of that search are used to create a list of accessible URLs--any pages shown as a result of that search will be placed on the list. Then, provided that there are more URLs on the list of URLs visited by the user in temporal order, the next URL on the list will be examined. (If there are no more URLs on that list, then the process repeats for the next user). If there is another URL on the list of URLs visited, then that URL is examined, and if that URL is on the list of accessible URLs, then

that URL is visited and the links on the page visited are added to the list of URLs, after which the examination of the user's list of URLs visited in order is continued. On the other hand, if the next URL was not on the list of accessible URLs, that indicates that the user has ended the search session, and the rest of the URLs visited in order is examined to find the next search.

Detailed Description Text (10):
In this way, each user's search activity is examined. The search session is considered active while the user is accessing pages retrieved from the search page, either directly or indirectly. A direct access from the search page is that access of a page listed in the search results; an indirect access is when the page is accessed from links appearing on a page which was either listed on the search page (a directly accessed page) or from links appearing on a page which was indirectly accessed. In other words, pages found on the list of accessible URLs which are then accessed are considered directly accessed, and pages added to the list of accessible URLs which are then accessed are considered indirectly accessed.

Detailed Description Text (11):
In another embodiment of the Shriver and Small method, several different searches may be considered active at one time. This can be useful because users may occasionally run more than one search at a time in different browser windows. In order to do this, a list of accessible URLs is maintained for each search, and instead of maintaining only one list of accessible URLs and comparing URLs on the users' list of URLs visited in temporal order with URLs in that list, several different lists of accessible URLs, each corresponding to a search performed by the user, are maintained.

Detailed Description Text (13):
In a preferred embodiment of the Shriver and Small method, the popular pick (or popular picks) for each search is determined. In this embodiment, the popular pick is defined as the last URL which has been reached as the result of a search, that is, the last link examined, and found to be on the list of accessible URLs. Since this is the last page that the user visited as the result of the search, it is assumed that that page was the goal of the user's search. This popular pick is stored along with the search in a database. In another embodiment, the popular pick may be the page or site which was explored for the longest time, or may be identified in some other way from the pages accessed by the user. In yet another embodiment, instead of just one popular pick, a number of pages may be stored as popular picks for one search. In another embodiment, users may designate pages as popular picks while performing searches, and pages so designated are stored, with searches, as popular picks.

Detailed Description Text (15):
After the creation of this database has been completed, and searches and popular picks have been stored, the preferred embodiment of the method examines users' HTTP requests in real-time, at the same time simultaneously sending the requests on to the web server to be processed. If the HTTP request is not a search engine query, the request is simply processed as per usual. However, if the HTTP request is a search engine query, the database is searched to see if the query can be found in searches which have been stored, together with their popular picks, in the database. If such queries are found in the database, the corresponding popular picks are displayed for the user along with the results from the search engine. The popular picks might also be displayed before the search engine results arrive at the web browser. If there are a number of popular picks, then they can be arranged in order of (presumed) importance to the user, e.g. in order of frequency--that is, if three searches result in two popular picks, the popular pick that was the result of two searches will be listed first, followed by the popular pick that was the result of only one search. To make the data more meaningful to the user, statistics such as the percent of the time that the popular pick was chosen when the query was searched can be displayed, along with the popular pick URLs.

Detailed Description Text (20):
In a preferred embodiment of the Shriver and Small method, the data is aged so that popular pick URLs from searches that are older than a certain number of days are assumed to be of lower importance, or may be purged from the database entirely. In still another embodiment of the invention, a periodic search of the popular picks is performed to ensure that they are still valid URLs. In yet another embodiment of the invention, when the user has input a search for which the database contains one or more popular picks, a request for one or more of these picks is sent to the web server. Then, in one alternate embodiment, the pages are presented to the

user, and in another alternate embodiment, the pages, having been requested, are more quickly available in a cache.

Detailed Description Text (21):
In a preferred embodiment of the Shriver and Small method, the real-time monitoring of users' request for search requests and provision of popular picks is implemented as an add-in module to the Apache web server (available from the Apache Group, www.apache.org).

Detailed Description Text (22):
An alternate embodiment of the Shriver and Small method uses this general technique of community based filtering to track URLs visited by members of the community, instead of only counting those visited as the result of a search. Each time a URL is, visited by a member of the community, that visit is logged in a database. A simple counter corresponding to that URL can be created (if no visits have been previously logged) or incremented (if the URL has been previously visited). This value of this counter can then be used as a measure of the importance of the corresponding URL to the members of the community. As above, the counter value can be used when deciding how to order or display search results.

Detailed Description Text (24):
According to the present invention, a search engine system will be able to identify more closely what the user intends when searching. This can be accomplished by initially pre-processing the search query and then using information about the user to present search results to the user in a manner that coincides with the user's preferences, interests, etc. One method is to provide feedback to the user while language is being parsed, disambiguating text by the use of lexical indexes, an expert database and data store of known information, formulation of known mental models for subject specific and area specific data. Characterization of results is based upon target user knowledge base as would be performed by a subject matter expert such as a librarian.

Detailed Description Text (25):
That is, when text is input on a form, or in a search input field, it can be sent to the search engine. If there are thousands of results, and the search server determines that the data returned are spread over a number of subjects, sources, etc., than the word or phrase typed into the input field of the search can be checked to see if it has multiple meanings, has special meanings in a dictionary of special word usage associated with a specialty or COI, or if the catalogue of expert knowledge has the word or phrase catalogued. The record in the COI or expert database could be used to select likely elements that would be of interest to the user.

Detailed Description Text (27):
In preferred embodiments, search results are presented to the user in the form of scatter grams and the like. The overall search results are displayed showing the relationships among resulting elements, with color, shape, and other attributes to differentiate among resulting elements being used. Preferably, icons and hot spots with text description are used to view details.

Detailed Description Text (28):
The benefits of the present invention can be realized in the following exemplary situation. An 8.sup.th grader searches for information about AIDS. A librarian would be able to identify many things about the person that will help select the right search results from among all the possible items returned. Things a librarian might observe about a person include items such as age, apparent level of education, context of the person (work, school, shopping), language skills used during the dialog (level of language, choice of jargon, language), etc. The librarian would be able to identify sources that are appropriate for the age, language skills, and scope of the search. A college graduate, on the other hand, may be looking for information on AIDS. The librarian would be able to identify characteristics of the college student that play an important part in figuring out which search results are relevant. Such items may be identified by a system, and may be more accurate over time. A search application system could function as a personal librarian who knows a lot about the person doing the search, and who can do a good job matching search results including sources, depth, breadth, age of items, and other features, with the user. The search engine would know a lot about a person from the collective knowledge of others in the COI, implied characteristics of the individual, and some items that can be assembled into a profile over time.

Detailed Description Text (30):
A web server front-end 18 is connected to the internet 16 and a server back-end 20 for the purpose of providing information search functionality to end user 12. Included in the front-end 18 are a communications interface 22 for sending and receiving message packets to and from end user 12, a page-by-page GUI sequence interface 24 for generating outgoing queries and receiving HTML pages, and a back-end interface 26 for communicating search information to and from back-end 20. It should be appreciated that while a GUI 24 is preferred, other interfaces such as audio interfaces could likewise be incorporated in the invention.

Detailed Description Text (42):
As an example, FIG. 6 shows a GUI architecture 202 for displaying search results when a large number of elements has been returned. GUI architecture 202 comprises a bird's eye view 203, a more detailed view 204, and a fully detailed view 205. The bird's eye view 203 is analogous to a geographic map showing a high level overview of the search results. The more detailed view 204 is analogous to zooming in to an intermediate level on a geographic map, and the fully detailed view 205 is analogous to zooming in to the most detailed view on a geographic map. View 205 will provide details of all data, means to retrieve pages such as links, etc. The user, of course, will only be presented with one of the above-mentioned views at a time, and will be able to select more detailed or less detailed views as needed.

Detailed Description Text (45):
Search query formulation function--Interactive form to obtain user requirements for the search including input fields, pull down windows with access to tools for formulating queries, configuration management window, etc. The user can identify types of queries to formulate and can select the functions from one of the tool windows. The configuration tool can add the individual functions together to provide a solution set matching the user's requirements. The configuration of the search can be accessed and can be modified at each node and the direction of the search may be changed on a subset of the results. A larger body of results can be obtained by selecting one of the nodes and delving deeper into a source, a feature, or another characteristic of the search results. The user can see that many of the results come from the magazine, "Living," and then decide to ask to view more of the "Living" references online.

Detailed Description Text (47):
Guide to reference materials--This guide is initially generated by provisioning the Source Characterization Database 52. The system continues to enhance relationships between resources and users as it learns from the process of users conducting searches and selecting items from among the search results. The users can then look at the selected resources typically used by member of the specified Communities of Interest. These resources can be presented in an order of most selected or another preferred arrangement.

Detailed Description Text (49):
Definition disambiguation function--A learning application is used to identify words or word groups with multiple meanings. The application first checks with the Communities of Interest in the Personality Profile of the user, and if there is a match between the subject of the definition and an area of interest among the users in that COI, then the search results are ordered with those items (COI relevant) and features displayed at the top or front of the results listing, table, or graphic. Other results which may not be a match with the COI indicators are still in the search results, but they are not featured at the top of the list, table, or graphic. The relationship among items is displayed as shown in FIG. 7(a) wherein, for example, disambiguation function window 206 is displayed to the user with COI clusters 207 displayed in the uppermost portion of the window. Alternate meanings of a word such as "lead" are displayed at the next lower level, such as a meaning 208 under the context of metals and meaning 209 under the context of electrical, etc. With the results of one COI clustering in an area and results of another COI clustering in another, the display method helps the user to see there are different types of answers to the search results, including but not limited to results from entirely different subject domains.

Detailed Description Text (53):
"Tell me more about this" function--Once search results have been returned, the application has more data about the search and its resulting elements. It knows what the sources are, etc. The user may wish to learn more about things the way they can in the library. Once a person gets to a bookshelf or magazine rack in a library, many items related to the original search are

present. These items may not have been explicitly listed in the search results but they are of potential interest to the user. The graphical user interface permits the user to continue a search by selecting one branch of the search and continuing in a more refined manner into the data that are related to the original search. Convenient methods of performing this function include using the Source Characterization Database 52 to categorize the results and preferably provide a link to results titled "Would you like more magazine articles about pedigreed cats?" The user could then access the information if desired. The Community of Interest is also used to prioritize the questions asked of the user. The Expert Knowledge Database 48 maintains and updates the query formulation function.

Detailed Description Text (54):
Prioritize features sub-function--The user has the opportunity to use an automated method of prioritizing features. This automated method relies on the expertise of the librarians who are informed about the subject matter, about searching, and about resources. These preferences are available in the Expert Knowledge Database. This database can learn and modify itself to more finely differentiate the resources used by the populations of users. The user can also select priorities by looking at the features presented on the graphical user interface including but not limited to: title, author, resource, date, persons, also about, etc. The third method of redefining a search by prioritizing features of the data includes selecting a Community of Interest profile to impose a priority on the search results.

Detailed Description Text (60):
The database will contain a blueprint to assist in identifying relevant items from a list of search results. The user can select one of the COI's from a list to self-identify. A provisioning process may assign a person to a COI (as in a company database, job skills database, information/corporate information directory, etc.). The database will be pre-populated with terms representing known features of interest to the COI. When search results cluster around these features, they will be used to determine display priorities over other features that may be identified in the data.

Detailed Description Text (66):
Creating a COI profile in profile catalogue 50 for a user with or without registration by user would be done by: assessing technical papers, web site, email, subscriptions (including source characteristics), mail lists to and from, items forwarded, organization charts, directory entries, query entries of the individual user 12 to determine special interests and clustering of features in the elements identified (a selection of these and other items can be assessed depending on access provided through group service arrangement, or individual subscription); borrowing a personality from a generic group similar to the individual then making changes as more information becomes available; assessing selected items in search results relating to an area of interest to determine characteristics and associating them with the user's profile; determining an age or social context of the individual; determining groups and individuals associated with the individual by reviewing email, technical papers, telephone and email address book, post, articles containing individual's name or member of group, etc.; determining an individual user's relationship to the group: casual, involved, etc.; recording multiple COI's of the individual using the system; determining attributes relating to understanding and use of language including but not limited to: primary language used, other languages used, competence and performance in each; and, determining attributes of language for each COI including but not limited to: access to vocabulary of each specialty, estimated sophistication in subject or area of interest, facility with vocabulary in specialty (competence and performance).

Detailed Description Text (86):
Results of a search should be processed by the search application to reveal to the user information about the items returned and also information revealed by the data returned. The resulting items should cluster in selected subspaces. The search application can use COI's of the user to disambiguate clusters in the elements returned. The search application can offer alternate views of the results based upon other known COI's that would use the data from sources returned. Data clusters in subspaces can be identified as significant based upon the interests of the COI group.

Detailed Description Text (87):
In this regard, Display Techniques Database 58 contains a database of forms, formats, charts, tables, graphics, color references, calendars, icons, etc. of known data display images, audio,

etc. It contains cross references for age, education, Community of is Interest. Level of language use, etc. to determine how to present search results, data elements, summaries, queries, alternative features (of data) list, optional functionalities of search applications, etc.

Detailed Description Text (96):
In addition to learning about the words used, the search results selected by other members of the same COI recently is used to order the search results for the current search as noted by Shriver and Small in the incorporated by referenced patent noted above.

Detailed Description Text (100):
Smart Search system 10 will query the user 12 for further input or selection from among a list of possible topics to resolve ambiguous search terms. This can be done either initially as the search terms are being collected in the user interface 24, or after an initial search has been performed, and a number of search results are analyzed by the Smart Search system 10 applications. In some cases it will be done automatically when the personality of the user 12 is understood and the ambiguity is mitigated from that understanding.

Detailed Description Text (102):
Smart Search system 10 will review the data returned from a search and identify sources, dates of publication, scope of results, scatter of results, domains of search results. All results from the search will be identified, however, by using the personality of the user 12 from personality profile 50, elements matching the likely concept, and breadth or depth of sources will be organized into a results presentation. The user 12 can always pass by the initial results presentation and see the complete list of items returned.

Detailed Description Text (103):
Search results sources will be catalogued to identify characteristics of the sources including but not limited to: dates and frequency, COI, COI's, general level, degree and level of jargon, categories of resources (calendars of events, classified, technology reviews, etc.) and will be used to inform user about sources. The source characteristics catalogue 52 will record whether the source is a primary source, secondary source, web site, commercial publication, technical, business, fiction, or other type of source. Experts will annotate the source. Source participants can add information to the record about the source.

Detailed Description Text (105):
The complete results presentation will contain representations of each distinct concept represented in the results and not constrained by COI, and a reference to each data point in each concept representation. Since many elements in a search result list belong to more than one category or abstract concept, the relationship among categories will be presented.

Detailed Description Text (106):
The user 12 can see a presentation that will indicate the scope and scatter of results for a given query. In this way, an individual user 12 can see what information was accessed from Information Database 42, and the source of the information, and learn about the age of materials accessed. The interface 24 can be configured to reveal resource management techniques to the user 12. By knowing about the source and COI preferences, Smart Search system 10 can teach the user 12 which sources are used to get information by a COI. So a person entering a new field of investigation can learn which resources would be ideal for a person in that field. He or she would not have to try to figure it out from the long list of search results. Again, the user 12 can prune the list of returned elements or expand what is included in the list by manipulating the data as it clusters into source, date, time, author, type and other characteristics of the data.

Detailed Description Text (113):
Search results may be presented to the user and the user may decide the results are not precisely what was sought. The user has changed the requirements in some way based on something learned from the first search results. The new search is not quite a new search, but a continuation of an existing search. This may be accomplished in a number of ways. In one example, the search can be modified by changing the choices of sources, or other attributes filtered by the policy representing the Community of Interest. These attributes, represented in the profile database, may be displayed to the user so some or all of them may be adjusted to more closely satisfy the search needs in the given instance.

Detailed Description Text (114):
Some search results may be out of line. One case when this might happen is when a person is doing a search for someone else. The search application may learn something inaccurate about the user. One feature of the search application would allow for the user to view the characteristics the search application has associated with him or her. For example, a window may be opened to show the COI's the user is a member of, and details of the COI's including age, education, geography, membership in clubs, etc. If the attributes are inaccurate, they may be changed or deleted.

Detailed Description Text (115):
Another means for altering the direction of the search application is to add clarification by looking at the data represented in the results. By selecting other features of the data to display relationships not clearly illustrated in the first display of the data, a search result may reveal new information to the user. The results data is representing the view from a particular point of view or COI. The user is not limited to this presentation of the data. Other features identified in the search are available and may be selected from a window, or by another means of manipulating the data of the results. A helpful device would be a window that lists all the recognized significant data relationships in the search results. For example, if 20% of the search results were from technical journals, and 15% of the search results were from items dating within the past ten years, those facts would be listed in the window for the user to see. The data itself reveals this information, using the database stored in the search application server to verify sources, publication dates, subjects and relevance, etc. Attributes about the users of the source are also relevant to the COI. If a description of the users of a source is made available, then the COI can be matched more directly with the targeted market of the sources. Even if the results are displayed by categories assumed to be important to the COI, there may be a feature in the data that is of interest to the user and is not graphically represented. By using a method such as the window mentioned above, a user may identify a feature that is of interest, and request that the display be modified to illustrate that feature.

Detailed Description Text (118):
Search Application: Prefers Theater Directions Magazine as a source for search conducted by a member of the Thespian COI. A search may result in a listing of results spanning 3 major unrelated COI's. For example, a search on Home Networking may result in thousands of items in a list spanning many pages of titles. When looking at the results, they may fall into the categories of: home communications network architectures including such subjects as computer LAN's, etc., home health aides visiting the ill or elderly, and home school programs using a network of teachers and other resources. By using the strategy associated with Communities of Interest, the system would identify which major subject area was related to a particular COI through: knowledge of the sources of materials, subject handled by the source, types of authors, types of content, disambiguating words used for the search and filtering for features related to the COI, etc. Without specific knowledge of the COI, the total of the search results would be segregated based upon strata and subject as defined in the global COI database. The results would then be displayed showing the features that differentiate them, in this case being home electronics/communications, social work, education, and other.

Detailed Description Text (121):
Humans are particularly good at identifying patterns in data when the data is presented in an appropriate graphical form. The key is to identify reasonable graphical representations for abstract data elements. It is another object of the present invention to identify subject matter and object classifications that will be modeled graphically so search results may be presented to the user 12. An example for representing solutions graphically would involve using a metaphor for placing pushpins representing search returned objects at a high level and with relationships described graphically (between the returned objects). Color can be used to indicate proximity characteristics. Sound can be used to illustrate relationships among items, and to represent navigational distances between objects. Sounds can enhance an application so audio-based interfaces may be used in addition to graphical. For example, a search might be conducted using a cell phone, and the user 12 will listen to results receiving audio cues to help formulate relationships among returned items. For example, suppose a consumer is searching for a particular car. The consumer knows some attributes relating to the car, such as the manufacturer and the year. The Smart Search will enable the consumer user 12 to use a configuration management approach to filling in a form through which the search will be

formulated. The assumption here is that the consumer's shopping interests can vary along a continuum from a very directed search through to a rather unconstrained browsing type of search. An important issue relating to these dimensions of searching is the ability for a search engine to have varying degrees of information presented during the query and still return useful, targeted, and understandable results.

Detailed Description Text (127):
The consumer may also wish to know how many of the results of the search are primary sources, and how many are duplicates of the same content or provider. The graphical display should provide a perspective to the consumer about how many unique hits the search has returned by evaluating certain attributes of the results. For example, journal, date, length of article, word comparisons on search results, URLs, etc. These relationships will be defined in the index database 42 and will be expanded on an ongoing basis to include more subjects, resources, sources, and other attributes.

CLAIMS:

1. A method for information searching comprising: determining at least one of interests and preferences for a user conducting an information search; conducting the information search; evaluating results of the search based on at least one of the interests and preferences; learning at least one of new interests and new preferences for the user based on the evaluating; manipulating the results based on the evaluating including: determining if the search returned in excess of a predetermined large number of results; and, clustering the results into feature groups based on results of the determining; and, summarizing the results into a display entity with dynamic attributes based on the manipulating.

11. A system adapted for information searching, the system comprising: a graphical user interface to facilitate communication between a user and the system; and search components operative to determine at least one of interests and preferences for a user conducting an information search, conduct the information searching, evaluate results of the search based on at least one of the interests and preferences, and manipulate the results based on the evaluating, including clustering the results into feature groups if the results exceed a predetermined number of results; a learning engine operative to learn at least one of new interests and new preferences based on the evaluated results; and, a display integrated into the interface to display search results based on the interest and preferences.

Previous Doc Next Doc Go to Doc#

L4: Entry 2 of 19 File: USPT May 11, 2004

DOCUMENT-IDENTIFIER: US 6734886 B1
TITLE: Method of customizing a browsing experience on a world-wide-web site

Brief Summary Text (7):
The WWW allows a server computer system (a Web server) to send graphical Web pages of information to a remote client computer system. The remote client computer system can then display the Web pages. Each Web page (or link) of the WWW is uniquely identifiable by a Uniform Resource Locator (URL). To view a specific Web page, a client computer system specifies the URL for that Web page in a request (e.g., a HyperText Transfer Protocol ("HTTP") request). The request is forwarded to the Web server that supports the Web page. When the Web server receives the request, it sends the Web page to the client computer system. When the client computer system receives the Web page, it typically displays the Web page using a browser. A browser is a special-purpose application program that effects the requesting of Web pages and the displaying of Web pages. Commercially available browsers include Microsoft Internet Explorer.TM. and Netscape Navigator.TM..

Drawing Description Text (5):
FIG. 2 is a schematic diagram of one of the many possible Web servers able to support the computational needs of the present invention;

Detailed Description Text (10):
Referring to FIG. 2, the Web site 10 is located on a Web server 54. A Web server is a computer that provides World Wide Web services on the Internet. Such a computer includes the necessary hardware, operating system, Web server software, communications protocols and the Web site content (Web pages) to provide the services. It will be understood by those of ordinary skill in the art that the particular details of the Web server 54 are largely irrelevant to the present invention. So long as the Web server 54 is capable of performing the functions described herein, any configuration will suffice.

Detailed Description Text (11):
As shown in FIG. 2, the Web server 54 includes a central processing unit (CPU) 56 for controlling the operation of the Web server 54, a hard disk drive 58 which includes the operating system 60, the necessary Web server software 62, the communications protocols 64, the Web site 10 content and a set of algorithms 66 for performing the various functions described herein, a random access memory (RAM) 68, a read only memory (ROM) 69, a floppy drive 70, a CD-ROM drive 72, an Internet interface 74 which connects the Web server 54 to the Internet, and a network interface 76 which connects the Web server 54 to the Operator's internal computer network. A similar configuration may be used for the other web servers discussed herein (i.e., the Registration Authority server and Certificate Authority server).

Detailed Description Text (12):
A typical configuration for the Web server 54 includes an Intel IP L440GX Motherboard; a Dual Intel PIII 500 mhz Processor; 256 mb 100 mhz SDRAM; 9.1 gb Ultra2/SCSI Hard Disk Drive; a Creative Labs 52.times.CD ROM; 3.5" 1.44 mb Floppy Disk Drive; Dual Intel PRO/100+Dual Port Server Adapters; Antec Rackmount ATX Case; Microsoft Windows NT Server V4.0 Service Pack 5; Microsoft IIS Option Pack v4; and Microsoft Data Access Components.

Detailed Description Text (22):
The indexed clinical data is then stored in the database 81 for reference and downloaded to the Operator's data warehouse server 88. After the data is stored in the data warehouse server 88, it is uploaded to the Web server 54 via the Operator's internal computer network and stored in the user database 44. The UAIs are encrypted by the Registration 15 Authority before being passed to the Web site (via the data staging computer 80) using first the Registration

Authority server's public key and then the date warehouse server's public key. As discussed further below, the UAIs are further encrypted on the user certificates on the Users' browser, and in the Certificate Authority's database using the Web site's public key. The purpose of this chain of encryptions is to prevent simple collusion among members of the operations staff at the various facilities (the Registration Authority, Web site, and Certificate Authority are housed at different facilities, and operated by different companies using different operations staff). The sequence of encryptions represents the order in which data moves through the system when first loaded. An upstream service can recover the UAI it provided, but the recipient cannot determine how the UAI was represented in the provider. Since both public and private keys are buried deeply in the code, only technical staff with sufficient time, opportunity, and skill to find and determine how to utilize the keys might be able to collude to compromise the system. This should be significantly beyond the technical ability of routine operations staff. As always, the senior technical operation staff present a security risk that must be managed by non-technical means such as incentives and penalties.

Detailed Description Text (24):
After uploading to the Web server 54, the clinical data is stored on the Web site 10 in the user database 44. The user database 44 is a table which lists each User by UAI and includes (but is not limited to) columns containing the User's Web ID, the User's customizations activation code, the User's ICD-9-CM, CPT-4, NDC and HCPCS J-code history, the User's content profile designations, and other de-identified information of interest to the Operator. The concept of content profiling will be discussed in detail below.

Detailed Description Text (28):
At that point the Web server 42 will pass the session to the Registration Authority by linking over the Internet to the Registration Authority server 82. In order to ascertain the UAI of the User, the Registration Authority will query the User for externally identifying information including, but not limited to, the User's name, social security number, date of birth, gender, medical plan ID, and the social security number of the principal account holder (in the case of a family). The User will input the requested information and transfer it over the Internet, in encrypted form, to the Registration Authority server 82. The Registration Authority server 82 will then check its database 86 and verify the information. If the information is incorrect, the process is terminated with a message referring the user to the appropriate contact at the user's Health Plan. If the information, however, is correct, the Web server 54 is instructed to create a Web ID for the User having UAI "x". The Web server will then ask the User to select a Web ID. The User's UAI and Web ID are then stored in the user database 44.

Detailed Description Text (29):
Next, the Registration Authority server 82 will generate a customizations activation code for the User, which will permit the User to customize his browsing on the Web site. Rather than transmitting this information to the User over the Internet, however, this information will be mailed to the User via U.S. mail 95 for added security. This mailing will not contain the User's Web ID or password, in case it is intercepted by a third party. Once the User receives the code, he/she may use the code to receive customized services on the Web site 10, as discussed below. The code is also transmitted to the Web server 54 via a secure transmission mechanism and stored in the user database 44.

Detailed Description Text (30):
After the Web ID is created, the Web server 54 links to a fourth party called a Certificate Authority in order to create the password. The Certificate Authority server 92 will first ask the User to select a password. After the password is selected by the User and transmitted to the Certificate Authority server 92, the Certificate Authority server 92 will ask the User to supply a familiar set of identifying questions and answers for use in future challenges. This is accomplished by use of encrypted forms passed through the Web server 54. The Certificate Authority server 92 finally generates a user certificate having an encrypted form of the User's password and UAI embedded therein and stores the user certificate 91 on the User's computer 90 through the Web server 54. The Certificate Authority also downloads to the User's computer 90 an applet 93 which will be used to verify the password on subsequent logins, as discussed below.

Detailed Description Text (31):
At this point, the User has a password and Web ID so that he/she can log on to the Web site 10. Those of ordinary skill in the art will appreciate that the only party which has both the

User's Web ID and password is the User. Thus, to obtain the User's Web ID and password, someone would have to hack into both the Web server 54 and the Certificate Authority server 92. And then to correlate the User's clinical data to the User's identity, the Registration Authority's server 82 would have to be hacked as well.

Detailed Description Text (36):
The applet 93 checks the password against the password embedded in the user certificate 91, generates an authentication code, and transmits the authentication code over the Internet to the Web server 54, which then passes the authentication code to the Certificate Authority server 92. If the authentication code returned by the applet 93 is negative, indicating an incorrect password, the Certificate Authority server 92 will attempt to authenticate the User by seeking answers to the identifying questions previously designated by the User. This is done through encrypted forms passed through the Web server 54. If it is not already present, the applet 93 is downloaded first so that the forms are secured from interpretation by the Web server 54 or any listening hackers. If the User successfully answers the identifying questions, the Certificate Authority authenticates the User and passes the session to the Web server 54. If, however, the User is unable to answer the identifying questions correctly, then the login process is terminated.

Detailed Description Text (38):
If the User has been authenticated by the Certificate Authority Server 92, the Certificate Authority Server 92 will so advise the Web server 54 and the User will be logged in to the Web site 10.

Detailed Description Text (39):
It should be noted that for ease of discussion, the Web server 54 is used as the login server. In practice, however, as those of ordinary skill in the art will appreciate, a separate login server may be used so that the Web server 54 is not unnecessarily tied up resulting in slow service to the Users.

Detailed Description Text (42):
Once the User has logged on the Web site 10, the User may freely navigate and search for information of interest to the User. However, the User may also seek to enjoy customized service. In order to receive customized service, the User will, upon login, elect to receive customized service by transmitting his/her customization activation code to the Web server 54. The Web server 54 will then check the user database 44 to make sure that the customizations activation code presented by the User is valid. Once customizations are activated, they will remain activated for all subsequent logins until they are de-activated by unchecking a personal customizations box. However, customizations may be reactivated just by clicking the box to add a check. Barring administrative interventions such as account recovery, the activation code only needs to be used once.

Detailed Description Text (43):
In order to provide customized services to the Users based on their respective medical histories, each Web page (or link) on the Web site is indexed by ICD-9-CM, CPT-4, NDC and HCPCS J-codes in addition to keyword. This indexing is realized through use of the link table 46. The link table 46 is a list of the URL of every Web page on the Web site 10 with corresponding keywords, ICD-9-CM, CPT-4, NDC and HCPCS J-codes in their respective columns. The codes are assigned to particular Web pages by trained medical professionals based on the content of the particular Web pages. The assignments are highly discretionary and will vary from medical professional to medical professional. It will be appreciated by those of ordinary skill in the art that the present invention is not limited to the foregoing coding systems, but may be used with any coding system.

Detailed Description Text (44):
Referring to FIG. 6, in accordance with one method of customization, the Web server 54 may be configured so that upon login to the Web site 10 by a User, the ICD-9-CM, CPT-4, NDC, and HCPCS J-code history of the User as found in the user database 44 is compared to the link table 46. For each Web page having an associated code that matches any of the User's codes, the Web page is suggested to the User for browsing. The benefits of this method of customization will be apparent to those of ordinary skill in the art. Rather than suggesting Web pages based on the browsing history of the User, Web pages are suggested to the User based on the User's unique medical history, resulting in a highly informed customization process. This method of

customization can also be accomplished through the search engine, as discussed below.

Detailed Description Text (67):
Referring to FIG. 9, in accordance with another method of customization, the tracking database 52 also makes it possible to customize based on where other Users with similar medical histories have browsed. Thus, Web pages which have been visited by users having identical or similar ICD-9-CM, CPT-4, NDC and/or HCPCS J-codes may be suggested to the User. For example, assume that a User has disease X which is indicated by ICD-9-CM code Y. Upon login, the Web server 54 may be configured to search the tracking database 52 and suggest Web pages to the User which were visited by other Users having ICD-9-CM code Y in their medical histories.

Detailed Description Text (69):
Content profiles are determined by trained medical professionals and are largely discretionary. Thus, the Web server 54 may be configured on log in to look up in the user database 44 all Users having the same content profile(s) as the logged on User. The Web server 54 then looks those users up in the tracking database 52 and suggests to the User the Web pages previously visited by those Users.

CLAIMS:

20. The database system for providing information to registered users recited in claim 16, further comprising a database server receiving the indexed de-identified personal data from the second database and providing a login for users and authenticating a user based on a user's assigned anonymous identifier and responding to queries from authenticated users to generate search results to queries which are customized based on the user's personal data while maintaining user privacy and confidentiality in the user's personal data.

22. The database system for providing information to registered users recited in claim 21, wherein the customization of -search results for a user is based on a user's medical data.

Previous Doc Next Doc Go to Doc#

L4: Entry 6 of 19 File: USPT Nov 4, 2003

DOCUMENT-IDENTIFIER: US 6643641 B1
TITLE: Web search engine with graphic snapshots

Abstract Text (1):
A search engine manages the indexing of web page contents and accepts user selection criteria to find and report hits that meet the search criteria. The inventive search engine has an associated crawler function wherein display images of the web pages are rendered and stored as snapshots, preferably when the pages are indexed. The search engine reports search results by composing an html page with links to the corresponding page hits and containing snapshot reduced size graphic images showing the web pages as they appeared when fetched and stored as snapshots.

Brief Summary Text (14):
The search engine operator can use various methods to find or select web page addresses that will be loaded and analyzed or indexed in building the database. The methods may be chosen to expand or to limit the number of web pages that the search engine will access. As a result, the results of searches vary among the different search engines.

Brief Summary Text (17):
Examples of search engines include Hotbot, AltaVista, Yahoo, NorthernLight, Excite, etc. In addition, there are some search engine portals that run the same user query through a plurality of other search engines. The search engine comprises a processor that maintains a web page which the user loads by aiming his browser at the search engine URL (e.g., Excite's URL is http://www.excite.com/). The received page (namely the processed version of the html source code that is displayed) typically includes one or more Common Gateway Interface (CGI) boxes or similar form processing means by which a user who wishes to make a search enters one or more letter strings as search criteria. Boolean combinations of two or more strings often can be included or will be implied if not stated. The criteria typically are construed met if the specified words or phrases are found anywhere in the html source code of the target pages when last indexed. This includes portions that are not displayed (e.g., meta-tags and comments). The criteria can specify attributes other than the presence anywhere of a certain text string. This may be helpful, for example, to limit search results to finding files of a certain type (e.g., with URLs linking to a certain file extension type to find a certain kind of media). The criteria can also bracket out files in a selected date window.

Brief Summary Text (18):
The search engine compares the criteria to available information for web pages and sends to the user a report identifying the web pages that meet the criteria. The report to the user is transmitted in html source code. To generate the report, the search engine finds URLs for the selected web pages and inserts a list of these URLs into a shell form (i.e., an "empty" html source code file). The shell form has text and formatting to display title headers, possibly also ad banners and similar information. The URL list that is produced is inserted into the html shell. Each URL is flagged in the html source as identifying an html link (href=[etc.]). Thus when the list is displayed by the users browser, the user can select among the results and point and click or similarly highlight and invoke the html link addressing the page that the search engine considered to meet the user's criteria. This then loads the html source code directly from the remote page that was selected and the browser displays the current contents of the referenced web page according the html source code found there at that time.

Brief Summary Text (21):
The typical search engine reports more to the search than the URLs of the indexed pages that meet the searcher's selection criteria. The URLs themselves, which are formatted as hypertext links in the search report, sometimes provide information as to whether or not a search hit is

pertinent to the user's desires. For example the domain name associated with the page may identify an owner known to be in a pertinent business, or on the contrary may show that the search result is plainly not relevant to the search. The search engine typically also stores and includes in the search report listing one or two of the first lines of the web page that is referenced, which frequently includes a title that may be helpful to show quickly whether the selected page is of interest. The search listing also may show the date at which the web page was last updated or the date that it was indexed.

Brief Summary Text (23):
It would be advantageous if the presentation of search results could be supplemented to more effectively assist a user running a search to quickly and meaningfully separate the pertinent and irrelevant results. However, such a capability will only be useful if it can be accomplished without unduly adding processing time and storage requirements to the steps involved in preparing database information for search and in presenting the results to the user.

Brief Summary Text (25):
It is an object of the invention to provide an abbreviated representation of searchable data files, in particular Internet/Intranet/Extranet html data pages, which represents their text and linked graphics in a visual snapshot form to supplement representations such as introductory text passages and URL addresses. It is a further object to collect and process the necessary information before conducting searches and to store a relatively small graphic file in association with the search database for representing each potential hit. The respective graphics file is reported to the user when a search results in a hit on the file, namely by inserting a hyperlink to the stored file in the search report sent to the user as the search results.

Brief Summary Text (29):
These and other objects are accomplished by the improved search engine of the invention, for managing user search and selection of data files stored at distributed systems coupled at network addresses. In particular the search engine is effective to improve searching of hypertext web pages on the Internet. The search engine has an associated web crawler operable to address and load successive web pages, and to index text data associated with the successive web pages. In this manner the search engine obtains parameter information such as words appearing in documents, word proximity and other information that can be used to distinguish at least groups of the web pages from one another when conducting a search. The web crawler stores the parameter information in a manner that cross references the paramater information with the associated web addresses or URLs of the web pages. The search engine accepts user-submitted search criteria and conducts a search or the parameter information to select the associated addresses of web pages that met all or part of the search criteria. The results can potentially be ranked, subdivided into categories and similarly handled according to known search engine operation. According to an inventive aspect, in conjunction with obtaining the parameter information for at least a subset of the web pages subject to search, the crawler renders a display image of the web page that is being indexed, and processes the image to provide a reduced size graphic image file corresponding to a static visual presentation of each of the indexed web pages. This graphic image file preferably is stored in a compressed graphic file format such as GIF, JPG, or a similar file, the file address or URL of which is stored and cross referenced to the criteria in the database that identifies the corresponding web page. When a search is conducted and results in a hit on a web page, its graphic snapshot is linked to the search results reported to the user. In a preferred embodiment, acceptance of the user search criteria and reporting of the results are handled by html page exchange communications between the search engine and the user. The search engine is accessed by the user and provides a form page having CGI boxes or the like for accepting text and/or other selections from the user. The search engine conducts a search which identifies one or more hits that are reported to the user by sending an html search results page. The search results page is composed by the search engine as a function of the search results and may contain no hits or a number of hits. Each of the hits is identified in the search results by the graphic snapshot, and preferably also by text information that reflects the content of the web page hit. Preferably, the search results page is composed to include a hypertext link to the URL address where the graphic snapshot file has been stored by the web-crawler/database/search-engine processes, for example by an IMG SRC=[path.backslash.filename] command inserted in html source code. As a result, the image file is loaded by the user's browser when processing the search results page, which generally occurs after the display of text has been accomplished.

Brief Summary Text (30):
As a result, the search results appearing on the user's browser include links to the web pages that were found to meet the criteria (hits), and also a snapshot graphic image of the way that the web page appeared when rendered at the time of indexing.

Brief Summary Text (32):
According to an inventive aspect, the graphic image file that is produced is not necessarily identical to the appearance of the page when ultimately loaded by the user after a search. In addition to the fact that the web page may have changed since it was rendered into the graphic file, the rendering is accomplished according to a predetermined display configuration of the crawler when rendered. Nevertheless, the graphic is a useful and very quick means for a user to sift through search results and determine immediately whether or not at least some of the hits bear further investigation.

Drawing Description Text (5):
FIG. 3 is a block diagram illustrating operation of the invention in connection with executing and reporting the results of searches.

Detailed Description Text (2):
According to the invention as generally shown in FIGS. 1-3, the reporting of search results by a search engine 20, is improved and facilitated by offering each searcher or user 30 a visual representation 35 of the web pages found to meet the user's search criteria submitted to the search engine. The invention is particularly applicable to an Internet search engine but can also be applied to other networks 50 where the search engine 20 is available for managing user search and selection of web pages or similar files, stored at distributed systems 52 coupled to the network. The web pages, which may be considered data files, are found at addresses to which the search engine can link to load the data files, for example being accessible using URL addressing of the pages as hypertext markup language (html), file transfer protocol (ftp), telnet or other such file types. The data files may have embedded links to other data file or to graphics or other media files. The search engine 20 of the invention accepts user queries that characterize files of interest, searches for the files and reports to each such user the results of the search including network addresses of the files found to at least partly meet the query, enabling the user to link directly to the files, and also a snapshot of how the file will appear according to the most recent rendering performed by the crawler of the search engine.

Detailed Description Text (6):
A block diagram showing an improved Internet search engine 20 according to the invention, for managing user search and selection web pages stored at distributed systems 52 coupled at network addresses to the Internet 50 or the like, is shown generally in FIG. 1. FIG. 2 illustrates a succession of method steps and/or programmed operations of the system for building and adding to or updating a database 62 of searchable information. FIG. 3 illustrates a method and apparatus for conducting searches by accepting user queries 54, conducting searches of the database 62 and reporting search results in the form of a composed search report 80 containing visual representations or snapshots 35 that depict a presentation of how the selected pages would have appeared according to a default display configuration at the time they were accessed by the crawler 60.

Detailed Description Text (9):
The web pages are generally maintained on web servers 52 (FIG. 1) that are "remote" from the querying user 30 and from the search engine 78, but actually could be anywhere that is addressable on the particular network, including on the user's own system. The web servers 52, in known manner, store text and graphic data or addresses of graphic data found elsewhere. That information is available upon request and in the case of the Internet and other TCP/IP protocol type networks is transmitted in packet form to any user that requests the web page by directing a request to the web server identifying the TCP/IP address of the web server 52, the sender's address or identity, and the address of the desired page. This normally involves addressing using URLs that identify the type of communication desired, such as transmission of an html page (versus a linked graphic or media file, or perhaps a different type of interface such as ftp or telnet), and an address that represents the domain name and a subdirectory path leading to the actual html file or other file.

Detailed Description Text (10):
The same sort of URL addressing is used internally in html pages to address image and other files that may be located at the same web server or elsewhere on the worldwide web, namely by providing a hyperlink that states the network address of the text or other content, as opposed to containing the content itself. Such hyperlinks can also be invoked to move around in a given file, for example from one subheading to another. The hyperlinks are embodied by automatically recognizable codes (e.g., "href=" or "img src=") that appear in the source code together with the various start and stop tags that specify text formatting, colors and other aspects of the page as it should be displayed, for example using a browser. In a browser such as MS Internet Explorer, the source of a displayed page can be displayed by selecting "View" and "Source" from the toolbar.

Detailed Description Text (11):
According to the invention, a crawler 60 collects web page data and is generally shown in FIG. 2. Crawler 60 can be operated preliminarily but preferably operates continuously during operation of the other components to collect additional data and/or to update data already collected. Crawler 60 has one or more fetching processes 66, several being shown in FIG. 1 and identified as Agent A (fetch) processes. The crawler 60 via its fetching processes 66 determines web pages to load and attempts to load them. For example, the crawler 60 may test TCP/IP addresses (known as scanning) or attempt to load pages from particular domain name addresses where servers might be up and running, obtained for example, from a domain name server (not shown). The text portion of any data obtained by the fetching processes 66 from a particular URL address is parsed or divided into discrete terms and statements. These terms and statements are compared to predetermined reserved terms and formats that represent URLs, file addresses and the like. When the comparison indicates that a hyperlink to another file or web server has been found (or that a given string so resembles a hyperlink as to be interpreted as such), the found address is added to a list of addresses and an attempt is made in due course to load a file at that address, thus increasing the field of files that have been consulted.

Detailed Description Text (12):
The general function of the Agent A fetching processes 66 is to obtain the files available from remote web servers 52 and to note the addresses of the files (URLs for the Internet) that when invoked will address and load the file. As a result of communication delays, it is preferred to employ a plurality of concurrently active requests for files so that one file can be processed while waiting to receive another. This aspect is represented in the drawing by plural Agent A processes 66, which obtain the fetched files and store at least part of the fetched files in a buffer memory or queue 92. In connection with html web pages, the data includes html source code, addressed files containing images, audio or other media, which are stored in buffer 92 together with the addresses from which they were obtained.

Detailed Description Text (14):
According to an inventive aspect, the crawler 60 that is operable to receive the web pages and to extract the parameter information from them, generates a file 72 of graphic image data corresponding to an appearance of each of the web pages, which is stored, preferably as a reduced-size and compressed image data file 75, in association with the database data respecting the page. When search results are reported to the user (FIG. 3), the search engine reports the associated URL addresses 82 of web pages that met the search criteria in a conventional manner, preferably inserting a hypertext link to each identified page into an html page reported to the user, optionally a short description or excerpt, and also inserts into the report page the graphic image snapshot file by inserting into the source of the report page a link to the stored compressed graphic image file 75. The user's browser displays the search results in conventional form, namely by showing a selectable hyperlink to the addresses and optionally a description or excerpt, and displays a snapshot of how the identified page is likely to appear if or when it is loaded by the user's browser, should the user point and click to the link to invoke the URL of the page hit.

Detailed Description Text (15):
The search portal 78 that performs the search by reference to the database 62 in storage media 64, reports the search by composing a web page containing the search results, assembling the search report using hypertext markup language. The search report contains headers and information identifying the portal and perhaps contains advertising. The search report also lists the hits that resulted from the search. More particularly, the search engine inserts (in list or table form) a text string showing the URL address of each web page hit (i.e., the pages

found to meet the user criteria) together with a hypertext linkage to that URL (e.g., an "href=" statement), causing the user's browser to show a link that can be invoked (pointed and clicked) to load the page at the stated address. Such a report is conventional in an html source search report. It typically also has a description or excerpt and may be arranged in a pyramid or hierarchy of categories. According to the foregoing inventive aspect, the search engine also inserts the URL address of the graphic file that has been processed by a further process identified in FIG. 2 as Web Agent B 95, to contain a snapshot reduced/compressed graphic 35 representing the page hit.

Detailed Description Text (16):
The link to the compressed rendered graphic file can be made, for example, by use of a IMG SRC=<domain>/<path><filename> command in the html source. The graphic can be associated with a hypertext link to the hit page URL as well as linking using an HREF=<URL of hit page> command as mentioned above. As a result, the user's browser when displaying the search results also displays the graphic snapshot image, as shown in FIG. 3.

Detailed Description Text (18):
Referring to FIG. 2., the search engine includes or is associated with web crawler 60, which is an engine that conducts web page addressing, loading and analyzing, and stores representative data in a storage device 64 containing a database 62. The stored representative data characterizes the web pages that the crawler loads and that are analyzed for content by process 68. Of the main activities to be effected by the search engine system (i.e., by the crawler and the search processor), preparation of database 62 allows a search to be conducted more quickly by reference to the processed database information gleaned from the field of possibly-selected files, than would be possible if the search engine attempted to load and analyze the entire universe of files after the user had submitted query 54 (FIG. 3), namely while the user was awaiting search results.

Detailed Description Text (29):
Deliberate as well as inadvertant "search engine corruption" sometimes occurs. It may be crucial for marketing or other purposes for a web site to be found in user searches on search engines, and it can be lucrative or otherwise beneficial for a web site operator if his/her site is ranked high in the search results for particular terms. Thus, a great number of website operators have ways to misrepresent the content of their pages. Keywords intended to cause the page to be selected and to rate highly in particular categories can be included and may or may not be displayed. Misleading text can be placed in miniscule font at the bottom of a page or misleading text can be hidden by making it the same color as the background on which it appears. Text can also be placed in "ALT" descriptions of images and graphics, thereby indexed by the crawler but not seen by the user. A particular term can be included one or many times to improve rankings, by one of the foregoing techniques, or by overloading keywords in "META" tags included in web pages and not displayed. Another technique is to temporarily post a page to be textually indexed by the crawler/search engine and then to replace its content after it has been indexed, or similarly, meta-refreshing the web page so as to redirect the user to another page address. According to an aspect of the present invention, the user can visually distinguish pages having undesired content and not waste time on them. Search engine corruption using the aforementioned techniques to provide misleading text is averted due to the visual nature of the present invention.

Detailed Description Text (31):
The snapshots 35 can be contained in formatted image files (e.g., GIF, JPG, etc.). The snapshot image files, or URL addresses pointing to the image files, preferably are stored in the database 62 that also contains the URL addresses of the indexed pages. In reporting search results, the search engine 78 inserts a link 82 aiming to the snapshot image file 35 into the html search results page 80. The search results appear on the users browser 84 as a link to selected pages with an associated snapshot of the page when indexed, as shown in FIG. 3.

Detailed Description Text (38):
The text data portion of a web page is most commonly five to ten Kbytes in length and is received in less than a second on a typical network connection. The text file is normally the first file sent from the originating web server. Image files and script or other code, if requested, follow afterwards. The robotic processes of requesting a text file, retrieving packets and reassembling the text file, parsing the text file by finding terms within delimiters, and indexing its contents, can be accomplished under normal circumstances in 0.5 to

1.5 seconds. Assuming a one second average processing time, one computer processor operating, for example, 25 text processing web crawler robots (which may be conservative), can obtain and index the text of 25 web pages per second every second. Operating continuously, such a crawler could process over 15 million web pages per week. Certain factors limit the rate at which pages can be processed. Web congestion, long files, long transmission sequences, low bandwidth server connections, and other factors that vary from one website to another and one time of day to another may limit processing speed. Nevertheless, a search engine portal that has several computers with multiple robots devoted to crawling the web, might complete an entire crawling sequence through a reasonable universe of selected web pages, in three or four weeks.

Detailed Description Text (52):
The search engine reports search results to the user that entered the search criteria, by composing an html source page and transmitting it to the user. This html report page may identify no hits or a long list of hits, depending on the search results. In composing the report page, the search engine typically shows the search criteria used, and displays indicia summarizing or similarly identifying each web page hit. For example, the search report can identify hits by the URL of the originating web page. Preferably a short text selection such as the first few lines of text is shown. The html coded report page prepared by the search engine includes an associated hyperlink to the URL of each hit. The URL can be shown in plain text and provided with an associated hypertext link (href=[URL]). The user reviews the URLs, sample text or other information and activates the hyperlink of a selected web page identified in the results, thereby loading the web page presently found at the address of the originating page when processed by the crawler robots.

Detailed Description Text (64):
The two general functions associated with preparing the database of information which is then subject to search and reporting, are the functions of retrieving all webpage data (performed by Web Agent A), and generating a "snapshot" file from the data (performed by Web Agent B). It is found that these functions can operate concurrently with or apart from the search engine processor or processors that search the database of information and return results to the requesting user. The preferred embodiment, however, is to perform all processing in regards to rendering, resizing, and compressing the snapshot prior to being accessible to surfers on the web. A cycle of processing (crawling, indexing, rendering) preferably is completed and the index and snapshot files that result are loaded into a database or are used to update a database, maintained on the server that accepts user search criteria and composes and sends to the user the search results.

Detailed Description Text (68):
In a preferred arrangement of the invention, the processing is accomplished in a network of programmed processors that are in a data communication with one another and each of which has a TCP/IP communication link to the web. The database containing the universe of crawled or to-be-crawled target web sites, which may number in the millions, can be stored in a controlling processor or can be part of a shared data store used to allocate individual URLs to client computers on the search system network, such as by permitting Web Agent A to obtain the next URL from the list and to flag the URL as in use. It is not strictly necessary to use the network paradigm. Instead, each Web Agent A or each client computer running multiple Web Agents of type A can contain its own database with a subset of the URLs of the universe, and the databases of a number of robots or clients can be synchronized periodically to eliminate duplicates, flag URLs after they have been crawled, and similarly updated. In a typical application, the database serves out a URL to the next Web Agent A in the queue and moves an index or "pointer" to refer to the next URL to be served out.

Detailed Description Text (72):
Frames also present a problem for the crawler robot regarding embedded html links to other web pages. The owner of a frames web page can include html links to web pages of others. If a surfing browser attempts to load the linked page by selecting (clicking on) the link on the frames web page, the browser will load the linked page but it will be within the frame of the first web page owner. The browser is not linked independently in that case and instead is linked through the frames page. Thus the html target address that appears in the browser toolbar and is recorded in the browser's history list is not a link to the selected site. Instead it is a link to the frames page, with a modifier that identifies the selected site. When that target address is invoked, the frame is loaded and the linked web page is inserted into the frame.

Detailed Description Text (79):
Animated GIFs and other changing features can also be identified by an icon indicating the presence of that feature. Preferably these animated features are selectively processed to provide a static image. Animated GIFs and some other technologies such as Macromedia Flash, provide an action sequence in the form of a plurality of images that are displayed in quick succession, normally in a loop. It is a problem with animations, especially those pertaining to Macromedia Flash Technology to select which frame will be captured or selected as representative of the animation. Animated GIFs begin with a graphic and the subsequent "frames" may be limited only to those pixels that have changed color from one frame to the next. Flash Technology usually begins with a blank screen or blank square. Choosing the first frame of a Flash movie as the designated frame to process and render would certainly be unaccepteable. According to alternative solutions, the Web Agent B can employ a timer to wait a predetermined time before capturing the rendered image in a file of the type that starts as a blank or fades in. It may be a matter of luck what in particular will be present at the moment captured in the changing portion of the display. An alternative is to generate a static image as a sum or average of two or more changing frames, which may produce a smeared static image. Another alternative is to disable the Flash plug in by a suitable message to the target site when loading the page. Disabling the Flash plug may eliminate any graphic data, namely if the website operators did not provide a static HTML page as an alternative to be presented for users who are not outfitted for Flash. Often, a user without Flash is presented with a blank screen with a tiny caption at the bottom reading "If you do not have Flash, click here." A rendering and subsequent snapshot of a screen similar to this could be misleading to the user if viewed within the search results of a search engine, so a timed capture is preferred.

Detailed Description Text (80):
It is an aspect of the current invention to provide an icon or similar indication within the search results as to whether or not a particular website contains Flash Technology. This alleviates possible inconsistencies in processing and rendering a Flash movie, and subsequent interpretation by the user of a search engine who may be viewing the snapshots. Moreover, for Flash and similar technologies that are optional for users, adding an indication of their presence benefits users of the search results. Specifically in the case of Flash, a user who has loaded the Flash plugin or otherwise has the capability to process the content will prefer to access pages that contain Flash content if other factors are equal. Users with browsers incapable of processing Flash technology might be forewarned that their browser may have difficultly rendering that particular website, or at the least would be neutral about that aspect of the web site. The use of Flash, RealAudio and other "value added" technologies is often an indication that a particular website has superior content.

Detailed Description Text (93):
When the user reviews the search report using a browser, the browser inserts the graphic snapshot image adjacent to the listing of the URL link to the subject web page. Thus the user can determine whether a page entry in the search results is of interest, not only from the text information included with the URL link such as a description and title, but also from a small size presentation of what the web page looked like when it was indexed.

Detailed Description Text (95):
There are some timing issues. Between the time that the web page was downloaded and the time that the user clicks on a search result entry to review the page, the contents of the page may have changed. If a website operator updated or changed the layout of that website since it was rendered and processed by the snapshot software (Web Agent A and Web Agent B), it is possible that the visual aspect as seen through the user's browser no longer coincides with the snapshot image in the search results. Nevertheless, the snapshot normally shows a mostly consistent visual representation of the current content of the web page.

Detailed Description Text (107):
In a preferred embodiment, the textual portion of search results always is sent and caused to appear first, prior to the snapshots corresponding to those results. As a result, regardless of whether the user has turned the snapshots capability "ON" or "OFF", the text portion appears first. If a user so desires, he can abort the transmission of the results based on review of the initially received portion. This is accomplished through programming within the snapshot server system that queues the text portion of the search results to be "released" or transmitted first, preferably even before addressing (or perhaps even checking for the presence

on the corresponding snapshots.

CLAIMS:

22. The improved Internet search engine of claim 21, wherein the search engine reports to the user the associated addresses of the web pages that met the search criteria, in a form of hypertext source data containing URL links to said web pages, and wherein the graphic image file is displayed in association with a URL link to the web page represented by the graphic image file.

Previous Doc Next Doc Go to Doc#

L4: Entry 9 of 19 File: USPT Jul 15, 2003

DOCUMENT-IDENTIFIER: US 6594694 B1
TITLE: System and method for near-uniform sampling of web page addresses

Brief Summary Text (4):
Documents on interconnected computer networks are typically stored on numerous host computers that are connected over the networks. For example, so-called "web pages" may be stored on the global computer network known as the Internet, which includes the world wide web. Web pages can also be stored on Intranets, which are typically private networks maintained by corporations, government entities, and other groups. Each web page, whether on the world wide web or an Intranet, has a distinct address called its uniform resource locator (URL), which at least in part identifies the location or host computer of the web page. Many of the documents on Intranets and the world wide web are written in standard document description languages (e.g., HTML, XML). Theses languages allow an author of a document to create hypertext links to other documents. Hypertext links allow a reader of a web page to quickly move to another web page by clicking on the links. These links are typically highlighted in the original web page. A web page containing hypertext links to other web pages generally refers to those pages by their URL's. Links in a web page may refer to web pages that are stored in the same or different host computers.

Brief Summary Text (7):
Referring to FIG. 1B, another way to obtain a random sample of URL's is to randomly select one or more search terms from a lexicon, perform a search engine query using the selected search terms, and then randomly select one or more URL's from the search results. The selected URL or URL's are added to the list of randomly selected URL's. This process may then be repeated until a suitably sized list of randomly selected URL's has been formed.

Brief Summary Text (8):
While the random sampling procedures described above result in a list of random URL's, the list is biased toward well connected URL's. Referring to FIG. 2, there is shown a small portion of a hypothetical set 50 of interlinked pages 51-65. As can be seen, some pages have only one inbound link, while others have much larger numbers of inbound links. The URL for a page that is referred to by many pages is more likely to be visited during a random walk, and also more likely to be indexed by a search engine than a URL that is referred to by few pages. Therefore the list generated by the aforementioned procedures is not uniformly representative of the URL's (or pages) in the set of reachable pages.

Detailed Description Text (3):
The Internet network 120 includes web servers 121 and a service known as a domain name system 122. It may also optionally include a web page indexing system 123. The web servers 121 store web pages. The domain name system 122 is a distributed database that provides the mapping between Internet Protocol (IP) addresses and host names. The domain name system 122 is a distributed system because no single site on the Internet has the domain name mapping information for all the web servers in the network. Each site participating in the domain name system 122 maintains its own database of information and runs a server program that other systems across the Intranet or Internet can query. The domain name system provides the protocol that allows clients and servers to communicate with each other. Any application may look up the IP address (or addresses) corresponding to a given host name or the host name corresponding to a given IP address in the domain name system 122. An application accesses the domain name system 122 through a resolver. The resolver contacts one or more name servers to perform a mapping of a host name to the corresponding IP address, or vice versa. A given host name may be associated with more than one IP address because an Intranet or Internet host may have multiple interfaces, with each interface of the host having a unique IP address. The domain name system 122 may be accessed by the web crawler 115 in the process of downloading web pages from the

world wide web.

Detailed Description Text (5):
The web crawler 115 includes a communications interface, or network connection, 102, one or more CPU's 101, an operator interface 103 (which may be remotely located on another computer), primary or main memory 104 and secondary (e.g. disk) memory 112. In an exemplary embodiment, the network connection 102 is able to handle overlapping communication requests. The memory 104 includes: a multitasking operating system 105; an Intranet/Internet access procedure 106 for fetching web pages as well as communicating with the domain name system 122; one or more threads 108 for downloading web pages from the servers 121, and processing the downloaded web pages; a main web crawler procedure, herein called the random walk module or procedure 110, executed by each of the threads 108; a set of URL's called the seed set 130; a list of visited URL's 132 that identifies the URL's of pages visited during a random walk; the list data structure preferably also stores the URL's of the outbound links in the visited pages (see FIG. 4); an unbiased sampling module or procedure 134 for sampling URL's from the list of visited URL's 132; a near-uniform list of URL's 135 generated by the unbiased sampling procedure 134.

Detailed Description Text (6):
Referring to FIG. 4, in a preferred embodiment the list of visited URL's 132 is stored in a data structure that includes a primary list of URL entries 136, each entry including a URL 137 and a pointer 138 to a list 139 of outbound link URL's. If a visited page contains no outbound links, the pointer 138 in the corresponding entry 136 is given a null value; otherwise it points to a list 139 of outbound link URL's stored by the random walk procedure. In embodiments in which the links between visited pages is not used by the unbiased sampling procedure 134, the pointers 138 and lists 139 need not be stored.

Detailed Description Text (7):
The URL's stored in the primary list of URL entries 136 are sampled by the unbiased sampling procedure 134. The URL's from the outbound links of visited pages are not sampled by the sampling procedure 134. Note that some of the unvisited URL's from outbound links may be invalid (there is no guarantee that the URL's in outbound links refer to existing data sets).

Detailed Description Text (8):
The set of addresses in the list of visited URL's 132 (i.e., the primary list of URL's actually visited by the random walk procedure) is sometimes herein called a set of randomly selected addresses, because the addresses in the list are generated by a random walk procedure.

Detailed Description Text (12):
The page at the selected URL (called the current URL) is downloaded (142). If the URL for the current page is not in the seed set, it is added to the seed set. Also, the current URL, along with the URL links in the downloaded page, are recorded in the list of visited addresses (144). The outbound (also called outgoing) URL links, if any, of the visited URL are preferably recorded (in an outbound link URL list 139, FIG. 4) by the random walk procedure so that this information does not need to be collected during computation of the page rank function, as described below. In embodiments where this outbound URL link information is not needed later, it is preferably not recorded in the list of visited URL's. If the downloaded page includes outbound links to other pages (146-Yes), a next URL is selected as follows. First, a random value r is generated (150). If r is less than a predefined value D (152), a next URL is selected at random from the seed set (140), and otherwise a next URL is selected at random from among the URL's in the outbound links of the current downloaded page. The predefined value D is preferably a value between 0.1 and 0.15. The comparison of random value r with predefined value D at step 152 is used to randomly limit the length of each walk (measured in terms of downloaded pages) that starts with a page in the seed set to a number whose average value is equal to the inverse of D. Thus, if D is equal to 0.14, then an average of approximately seven (i.e., 1/0.14) pages are downloaded by the random walk starting at each selected page in the seed set.

Detailed Description Text (18):
Once the random walk is completed, the sampling procedure is used to generate an unbiased sampling of URL's from the list of visited URL's. As indicated earlier, the list of visited URL's is heavily biased toward well connected pages that are referenced by many other pages. The purpose of the sampling procedure is to generate a list of samples of URL's that compensates for that bias, so as to generate a near-uniform set of URL's. It may be noted that

this list of samples, sometimes herein called a "set of samples" or a "sampled set," may include more than one occurrence of some URL's.

Detailed Description Text (19):
Referring to FIG. 6, the unbiased sampling procedure begins by computing a reachability measure or value for each URL in the list of visited URL's (170). There are at least two ways to compute or generate a reachability value. Referring to FIG. 7, the simpler of the two is to set the reachability value to a "visited ratio" (VR), computed for each URL u as follows: ##EQU1##

Detailed Description Text (20):
where u represents one of the URL's in the list of visited URL's. A page is "visited" during the random walk when it is downloaded (see step 142). "Visiting" a URL and visiting the page referenced by a URL are considered to mean the same thing.

Detailed Description Text (21):
Computation of the visited ratio requires that the random walk procedure keep track of (A) the number of times each URL in the list of visited URL's is visited during the random walk, and (B) the total number of page visits during the random walk.

Detailed Description Text (23):
where u is one of the URL's in the list of visited URL's, D is the aforementioned parameter preferably having a value between 0.1 and 0.15, N is the total number of distinct (unique) URL's in list of visited URL's, Out(u.sub.i) is the number of outbound links to other pages in the page at u.sub.i, and each of the k pages at URL's u.sub.i, for i=1 to k, is a predecessor page having a link to page u.

Detailed Description Text (24):
In order to be able to compute the M'(u) values, the procedure computing the page rank function will need to access the outbound link information stored in the list of visited pages by the random walk procedure. This outbound link information is used when computing the value of M'(u) for each distinct URL in the list of visited URL's in order to determine which pages have links to each page u for which a page rank is being computed. In other words, inbound link information for each page u is extracted from the outbound link information stored in the visited list.

Detailed Description Text (27):
The M'(u) values computed in 196 are normalized by summing the M'(u) values for all the distinct URL's in the list of visited URL's so as to generate a value X, and then dividing the M'(u) values by X to compute an updated M(u) value for each URL u (197). In addition, a root mean square error value RMS is computed to determine the "distance" between the previous set of M(u) values (if any) and the current set of M(u) values in accordance with the standard root mean square formula: ##EQU4##

Detailed Description Text (28):
The page rank function is repeated (steps 196, 197) until the error value RMS falls below a preselected threshold (198). The final set of normalized page rank values M(u) computed by the page rank function are used by the sampling procedure as the reachability values for the URL's in the list of visited URL's.

Detailed Description Text (29):
Once the reachability values for the URL's in the list of visited URL's have been generated (170), the sampling procedure (FIG. 6) continues. As indicated earlier, unique identifiers, such as 1 to N, are assigned to each of the distinct URL's in the list of visited URL's (172). A probability density function PDF(i) is computed (174) for each of the URL's, by computing the inverse of the reachability value M(u) for each of the URL's: ##EQU5##

Detailed Description Text (33):
Sampling of CDF, and thus of the URL's in the list of visited URL's is accomplished as follows. A random value r is generated, having a value between 0 and 1. Using the value of r, an index i is selected such that CDF(i-1)<r.ltoreq.CDF(i) (182), and then the corresponding URL (i.e., the one assigned identifier i in step 172) is added to the list of uniformly sampled URL's (184). Steps 180 to 184 are repeated until a termination condition is reached (186), such as when the uniformly sampled list contains a predefined number of URL's. The resulting uniformly sampled

list 135 (FIG. 3) can then be used for any of the tasks mentioned above in the Background section of this document.

Previous Doc Next Doc Go to Doc#

Freeform Search

Database:
- US Pre-Grant Publication Full-Text Database
- **US Patents Full-Text Database**
- US OCR Full-Text Database
- EPO Abstracts Database
- JPO Abstracts Database
- Derwent World Patents Index
- IBM Technical Disclosure Bulletins

Term: L3 and (list near url)

Display: 50 **Documents in Display Format:** FRO **Starting with Number** 1

Generate: ○ **Hit List** ◉ **Hit Count** ○ **Side by Side** ○ **Image**

Search | Clear | Interrupt

Search History

DATE: Friday, December 17, 2004 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=USPT; PLUR=YES; OP=OR			
L4	L3 and (list near url)	19	L4
L3	L2 and (search near result)	102	L3
L2	L1 and (link near (web near page))	545	L2
L1	web near server	7379	L1

END OF SEARCH HISTORY

Freeform Search

Database:
- US Pre-Grant Publication Full-Text Database
- **US Patents Full-Text Database**
- US OCR Full-Text Database
- EPO Abstracts Database
- JPO Abstracts Database
- Derwent World Patents Index
- IBM Technical Disclosure Bulletins

Term: L3 and (list near url)

Display: 50 **Documents in Display Format:** FRO **Starting with Number** 1

Generate: ○ **Hit List** ◉ **Hit Count** ○ **Side by Side** ○ **Image**

Search | Clear | Interrupt

Search History

DATE: Friday, December 17, 2004 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=USPT; PLUR=YES; OP=OR			
L4	L3 and (list near url)	19	L4
L3	L2 and (search near result)	102	L3
L2	L1 and (link near (web near page))	545	L2
L1	web near server	7379	L1

END OF SEARCH HISTORY